1C2-OS4b-7

自己組織化マップ SOM を用いた擬情語の分類比較

～ 確率的潜在意味解析 pLSA による効果の検討 ～

Psychomime Classification by Using Self-Organizing Map and Probabilistic Latent Semantic Analysis

黒澤 義明 [*1]
Yoshiaki KUROSAWA

竹澤 寿幸 [*1]
Toshiyuki TAKEZAWA

[*1] 広島市立大学大学院 情報科学研究科
Hiroshima City University, Faculty of Information Sciences

Our purpose of this paper is to automatically classify Japanese psychomimes, a kind of onomatopoeia, that represent users' feelings or emotional aspects, such as "pokapoka" and "tikutiku," by using a probabilistic latent semantic analysis (pLSA) and a self-organizing map (SOM) algorithm. Dealing with psychomimes is very important because they reflect the speaker's emotion and frequently occur in our daily communication in Japanese. However, it is difficult to communicate the meanings of them to people who do not understand them because they are not observed directly. We experimentally classify and visualized psychomimes with two methods, pLSA and SOM. Results demonstrate that our method is effective because the classification groups could be successfully depicted as a map by combining the methods.

1. はじめに

オノマトペとは擬音語・擬態語の総称であり，物事を的確に表現することを可能とする言語表現である．しかし，オノマトペは感覚的な表現であるため，その語義が曖昧であり，コンピュータにとって扱いにくい対象であり，的確な意味を理解させることは困難である．

オノマトペの中でも擬情語は，さらに伝達の困難を伴う．擬情語が人間の感情や情緒に基づく語であり，音由来や様態由来ではないため，擬音語のように聞かせたり，擬態語のように見せたりすることによって伝達することができないからである．このため，コンピュータに理解させるばかりでなく，人間同士の間でも伝達が困難となりうる．

この様な理解困難さを解消し，視覚化により意味伝達を容易にするため，我々は自己組織化マップ SOM を用いた，オノマトペの自動分類・視覚化システムを提案してきた [Kurosawa 10, 黒澤 10, 中村 09]．今回は新たに確率的潜在意味解析 pLSA を分類過程に導入し比較検討を行うことにより，新たな知見の獲得を目的とする．

2. オノマトペ分類と擬情語

オノマトペとは擬音語・擬態語の総称であり，物事を的確に表現することを可能とする言語表現である．日本人にとって，非常に重要な役割を担っていることは間違いない．このため，オノマトペは表現が非常に多い．辞書の見出し語として収録されている語だけでも 4000 を超える [小野 2007]．さらに，新語が絶えず作られるため，辞書に載っていない語も多く存在すると考えられる．

同様に，種類についても様々である．例えば，次のような分類が挙げられる．

- 擬音語
- 擬態語
- 擬情語
- . . .

連絡先：黒澤 義明，広島市立大学大学院 情報科学研究科，広島市安佐南区大塚東 3-4-1，082(830)1581，kurosawa@ls.info.hiroshima-cu.ac.jp

擬音語は，音に由来する表現であり，物理的な音を言語で表現しようと試みた結果である．『わんわん（イヌの鳴き声）』等がこれに当たる．

音に由来するため，意思の疎通が図れない場合，音を聞かせることで伝達が可能となる．例えば，日本語学習者に『わんわん』が伝わらなければ，イヌの鳴き声を聞かせればよい．おそらく，『Bow-Wow』のことと気づくであろう．

次の擬態語は様態に由来する表現である．例えば，『ぶかぶか』等が挙げられる．

擬態語の一部は，擬音語同様，人間の知覚に訴えることによって，意思の伝達が可能となる．例えば，先の『ぶかぶか』の場合，日本語学習者に伝わらなければ，大きなズボンを持ってきて履いて，見せてみればよい．おそらく，『baggy』と言う意味だと判断する．

しかしながら，このような擬音語や擬態語と異なり，擬情語については，知覚を通じて意思の疎通を図ることは難しい．個人の感覚・感情・情緒に根ざした表現だからである．例えば，痛みとともに生起する「きりきり」という単語の意味が簡単に伝えられるだろうか？ 聞かせたり，見せたりすることができない以上，言語表現を使って説明することになる．しかし，この説明は日本人にとっても容易ではないと考えられる．

擬情語をわかりやすく伝えるため，擬情語の視覚化が必要である．そこで，本研究は自己組織化マップ SOM を用いた先行研究 [Kurosawa 10, 黒澤 10, 中村 09] に，さらに，確率的潜在意味解析 pLSA 手法を導入し，擬情語の視覚化を試みる．

3. pLSA と SOM による分類と可視化

本研究は，擬情語の分類・可視化のため pLSA と SOM を使用する．ここでは，両手法についての説明を行う．

3.1 確率的潜在意味解析 pLSA

pLSA とは確率的潜在意味解析（Probabilistic Latent Semantic Analysis）のことであり，次元圧縮を確率的に行う手法である[Hofmann 99]．

pLSA では潜在変数 $z \in Z$ を導入し，文書 d における単語 w の生起確率を次に示すように定式化する．

$$P(d,w) = \sum_{z \in Z} P(z)P(d \mid z)P(w \mid z)$$

そして，潜在変数モデルにおける最尤推定のため，EM アルゴリズムを用いる．まず，E ステップを示す．

$$P(z \mid d,w) = \frac{P(z)P(d \mid z)P(w \mid z)}{\sum_{z' \in Z} P(z')P(d \mid z')P(w \mid z')}$$

次に M ステップを行う．なお，式中，*n(d, w)*は文書 d における単語 w の出現回数である．

$$P(w \mid z) \propto \sum_{d \in D} n(d,w)P(z \mid d,w)$$

$$P(d \mid z) \propto \sum_{w \in W} n(d,w)P(z \mid d,w)$$

$$P(z) \propto \sum_{d \in D} \sum_{w \in W} n(d,w)P(z \mid d,w)$$

pLSA は E, M 両ステップの反復を行い，生起確率が最大化となるモデルを生成する．この結果，文書内の単語が潜在変数を媒介とする確率として表現され，文書表現に必要とされる単語数（次元数）が，用意した潜在変数の数に縮約されることとなる．

ここで，同一潜在変数の寄与を有する複数の語は，同種の語であると考えられる．このため，pLSA はクラスタリング手法と捉えることも可能である．このとき，各クラスタへの帰属が，各潜在変数への確率値として表現される．複数の潜在変数が，特定の一語に対して，高い値を持つことも考えられるため，ソフトクラスタリングの手法として位置づけられることもある．単独のクラスタに分類を行うハードクラスタ手法と比べ，分類結果がわかりづらいという欠点がある．特に，次元数が多いと，クラスタリング結果を直感的に理解することは難しい．

3.2 自己組織化マップ SOM

本研究では，擬情語の分類・視覚化のため自己組織化マップ(Self-Organizing Map, SOM)を使用する [Kohonen 01]．SOM は，多次元ベクトルデータをその特徴を残したまま，2 次元マップに写像する．特に非線形のデータに対し有効であり，による擬情語の分類等，自然言語処理での有効性が確かめられている [Kurosawa 10, 黒澤 08]．

(1) 自己組織化マップのアルゴリズム

SOM は二層からなる神経回路網モデルである．入力層への入力により，競合層の特定の領域が反応するような，教師なし学習を行う．

入力層への n 次元の入力ベクトル x は，$x = \{x_1, x_2, \cdots, x_n\}$ と表現する．また，競合層にはノードと呼ばれるユニットがあり，全ノードから，入力層との間に参照ベクトル m と呼ばれるリンクが行われる（図 1）．

ここで，次式を満たす勝者ノード c の発見を試みる．次式は入力ベクトルに最も類似した参照ベクトルを持つノードを見つける操作と考えられる．

$$\forall i, \quad \|x - m_c\| \le \|x - m_i\|$$

勝者ノードの発見に続いて，近傍 $h_{ci}(t)$を決める．本研究では時間 t とともに減少するガウス関数を用いた．この近傍内では，複数の参照ベクトルを入力ベクトルに近づける操作を行う（図 2）．つまり，時間が経つにつれ，近隣のノードの類似性が増し，隣接ノード間距離が近づく（図 3の右中央部の変化）．次に，時間 t を用いた更新式を示す．

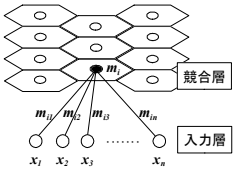

図 1　SOM の基本概念

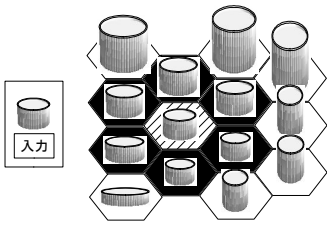

図 2　勝者ノード，近傍

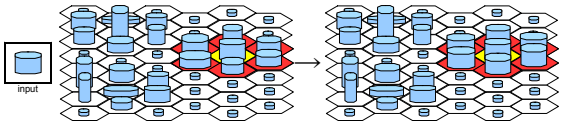

図 3　参照ベクトルの更新

$\forall i \in N_c(t)$を満たすとき，

$$m_i(t+1) = m_i(t) + h_{ci}(t)(x(t) - m_i(t))$$

それ以外のとき，$m_i(t+1) = m_i(t)$

以上の手続き～勝者ノード発見，近傍更新～を繰り返すことにより，似たベクトルが近くに配置されるよう，自動的に教師なし学習を行う．これが SOM のアルゴリズムである．

4. 実験

次に，pLSA, SOM を用いた，擬情語分類・視覚化実験について述べる．

4.1 実験で用いる擬情語

本稿で分類のために使用する擬情語は，我々の一連の研究 [Kurosawa 10, 黒澤 10, 中村 09] 同様，全部で 119 種類（表 1）であり，先行研究 [Akita 06] の定義を基にしている．彼の研究では，対象語にそれぞれ以下のようにシンプルな英語で意味が示されている．本研究では，この説明文中の特徴語（以下の事例中，イタリックで記述）に着目をし，特徴の分類を行った．この特徴をクラスタリングの際の正解例とする．

- dokidoki　‘feeling one’s ***heart*** throbbing’
- hinyari　‘feeling pleasantly ***cool***’

表 1　使用する擬情語

特徴	個数	着目した単語
目眩	9	eye, dizzy
鼓動	9	throb, heart, jump
痛み	9	sore, pain, …
臭い	5	smell
温度	14	cool, cold, hot, warm, …
刺激	9	pungent, skin
他	64	

例えば，表 1の「温度」では，cool や cold を定義に持つ擬情語が 14 個用意されていることを意味している．なお，定義内に共通の単語を持たない擬情語については，すべて「その他」を正解とすることとした．この手続きは客観的な分類を行い，適切な正解例とするためである．

4.2 実験手続き

実験手続きについて述べる. pLSA, SOM のどちらも多次元ベクトルを入力とするため, まず, その変換が必要となる.

(1) 擬情語の多次元化

本研究では pLSA, SOM による分類評価を行うための言語材料として, Web 日本語 N グラムを使用した [工藤 07].

先述の擬情語の直後に出現する動詞を抽出した結果, 1164 個の動詞と共起したため, 1164 次元のベクトルデータとして扱うこととする. ただし, 複数のベクトルの向きが同じでも, 極端に頻度に差があるときには, そのベクトルが示す語彙同士を異なる性質を持つと誤ってみなす可能性がある. そこで, 以下の手続きにより, 擬情語ごと動詞 v_i の出現率を求め, 各ベクトルの値とした.

$$v_i\text{の出現率} = v_i \Big/ \sum_{i=1}^{n} v_i$$

(2) pLSA 実行

上記の 1164 次元化されたデータを使用し, 工藤の実装[工藤]を用い, pLSA による次元縮約を行った. 今回は潜在変数の数, すなわち縮約後の次元数として 20 を指定した. なお, この値は便宜的に設定した値であり, 重要な意味を持つわけではない.

(3) SOM 実行

3.2で説明した手続きにより, som_pak を使用した 2 段階の分類学習を行った. 第一段階で広範囲に, 第二段階でより詳細に学習を行うためである. なお, 初期学習率係数 α, 初期近傍半径 r 等のパラメータは予備実験により決定された. 設定値を以下に示す.

マップサイズ:64 ノード×48 ノード

1st: 学習回数 1,000,000, α=0.05, r=80

2nd: 学習回数 10,000,000, α=0.01, r=40

4.3 実験結果と考察

上記の手続きにより得られた実験結果について述べる. pLSA 手法によって得られた結果, SOM により得られた結果をそれぞれ載せ, 最後に考察を行う.

(1) pLSA 実行後

pLSA を実行し, 前述のように 20 次元に圧縮した結果の一部を記す(表 2). 3.1に述べたように, 圧縮結果はクラスタと同室し可能である. なお, 本稿では特に,「温度」に分類される擬情語の分類結果を中心に考察を行うこととする.

表 2 pLSA による分類結果（一部）

擬情語	分類	確率
デレッ	他	1.000
ポカポカ	温	0.999
ぽかぽか	温	0.987
ほくほく	温	0.909
ジン	痛	0.520
ホカホカ	温	0.364
ホクホク	温	0.083

擬情語	分類	確率
いそいそ	他	1.000
ほかほか	温	1.000
シンミリ	他	0.283
ほっと	他	0.207
フラフラ	目	0.026
うっとり	他	0.025
ホカホカ	温	0.022
ぽかぽか	温	0.013
しんみり	他	0.010

擬情語	分類	確率
ゾクゾク	温	0.975
ガンガン	痛	0.971
がんがん	痛	0.947
マッタリ	他	0.913
ヒンヤリ	温	0.818
フラフラ	目	0.795
ほっと	他	0.768
ひんやり	温	0.089
ぞくぞく	温	0.026
ぷりぷり	他	0.012
おちおち	他	0.007
チクチク	刺	0.005

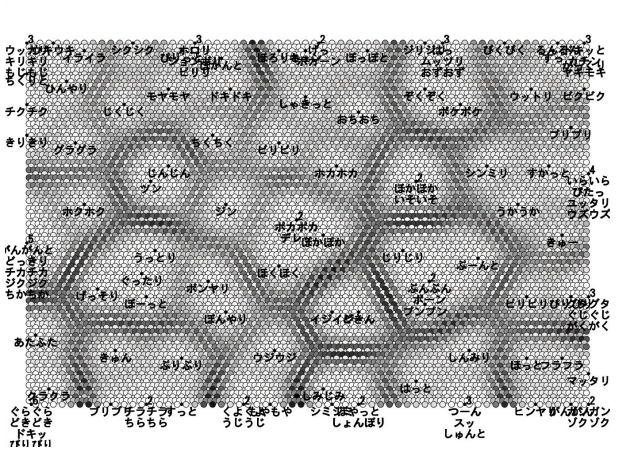

図 4 SOM による擬情語の分類結果

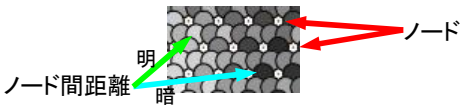

図 5 マップの見方

表 2の左側(赤)では,「その他」に分類される筈の『デレッ』や「温度」に分類される筈の『ポカポカ』等の単語の帰属確率が高いことがわかる. また, 7 個の単語のうち 5 個, すなわち, 約 70%が「温度」を正解とする擬情語であることから, このクラスタは温度に関連していると考えられる.

その一方で, 表 1に記したように,「温度」に含まれる擬情語は 14 語ある. したがって, 多くの擬情語の抽出に失敗していることがわかる(再現率, 36%).

また, 表 2の中央(緑)や右(青)では,「温度」関連の擬情語が存在してはいるものの, クラスタの特徴として表現されるまでには至っていない. pLSA 単独の分類が有効とは, 現時点では言えない. 今回は, 分類数を 20 個固定で行っている. 実際には変更可能であり, 変更に伴い結果が変わることが予想される. 現状では, 検討が不十分と言えよう. 今後, 詳細な実験が必要である.

(2) SOM 実行後1

次に, 上記の 20 次元に縮約されたデータに対し, SOM による分類を試みた結果を図 4 に示す. 図 4では, 隣接ノード間距離の最大値と最小値を元に, ノード間の距離が 0-1 になるよう変換し, 明度で表現した図である. また, 図 5 にマップの見方を示す.

赤線の先に示した小丸がノードを示す. また, 図 5 中, 扇のような形状により, ノード間距離を示す. 白(緑線)が近く, 黒が(水色線)遠いことを示しており, グラデーションにより表現した. 暗い輪郭を持ち, かつ明るい内部を持つ領域は, 外側とは異なる特徴を持っていると考えられる.

図 4の中央部に『ポカポカ』『ほかほか』等の「温度」に分類される語が集まっていることがわかる. この点については, 前項の結果同様, 抽出漏れがあるものの, 抽出そのものは評価できるという結果になっている.

ただし，指標として採用される精度・再現率については，先行研究 [Kurosawa 10, 黒澤 10, 中村 09] で行っているような領域の定義なしに，もとめることはできない．今回は紙面の制約もあるため，精度等の議論は省き，今後の課題とする．

(3) SOM 実行後2～特定の変数のみに着目

pLSA による縮約結果のうち，「どの変数が結果に影響を与えているか」をわかりやすくすることを考える．このために，図 4の表現を変更し，特定の変数のみに着色を行った結果を図 6に示す．ここで着色した赤・緑・青は，表 2の変数に対応する．色が濃いほど密集しており，薄いほど遠く離れていることを示す．なお，四角で囲まれた語は「温度」グループに属する擬情語である．

ここで注目すべきは，緑で着色された変数である．表 2に示したpLSA の結果では 9 個の擬情語が含まれていたのに対し，図 6では 4 個だけになっている．すなわち，『いそいそ』，『ほかほか』が中央に，後は色が薄い『シンミリ』が周辺に，『ホカホカ』が赤との境界（「赤色＋緑色」で表現）に配置されているわけである．これ以外は他の変数の影響を受け，緑の範囲から除外されたことになる．つまり，pLSA の解釈が難しいという欠点を補い，クラスタの整理が可能であることを意味している．

さらに，赤と緑の関係を考えれば，両者は『ホカホカ』という結合点で結びついていることがわかる．このとき，色が薄い『ジン』や『シンミリ』は，この結合点とは反対の方向に配置されている．したがって，適切な閾値を設定することにより，「温度」グループから除外できる可能性，そして，新しいクラスタが構成される可能性がある．ただ，前項でも述べたように，領域の定義なしに，議論することは容易ではない．別の機会に議論したいと思う，

(4) 問題点

擬情語の定義に関する問題点を述べる．

表 2の右（青）と図 6の右下（青）のグループに擬情語『ガンガン』『がんがん』が入っている．「冷房をがんがん冷やす」等の例文から「温度」という分類がされたと考えられる．周りに『ヒンヤリ』と『ゾクゾク』があるため，特に低い温度を示していると思われる．この意味では正しい実験結果と解釈してもよい．しかしながら，正解例を作る際に『がんがん』は「頭ががんがん痛む」の意の「痛み」に含まれているため，結果として，不正解扱いとなっている．

本研究においては，擬情語に複数の語義が合ったとしても，語義は 1 つしかないとして正解を作成している．本来，擬情語を含むオノマトペは複数の解釈が可能であるため，正解の作り方については再考する必要がある．

同様に定義の問題で言えば，「その他」に含まれる擬情語が多いことが問題である．今回の実験結果（図 4）を見ると，中央下部に『くよくよ』のグループと『しみじみ』のグループの 2 種の異なるグループがあることがわかる．これらのグループは人間の心情を表す擬情語が多く含まれており，両者の違いを含め，考察する価値があると思われる．しかしながら，正解作成の論文 [Akita 06] の定義中に共通の単語がなかったため，「その他」に含まれることとなった．

正解の客観性を保つための操作であるとは言え，重要な考察が行えないのは問題である．他の客観的な正解作成手法を考えた上で，再度実験を行うべきであろう．

5. おわりに

本研究は擬情語の分類・視覚化を目的とし，pLSA と SOM を用いた検討を行った．pLSA と SOM を組み合わせた結果，pLSA による分類だけでは明確でなかった分類を，新たに見つけ出す可能性が示唆された．この点，本研究の提案手法の有効性が確認されたと言える．

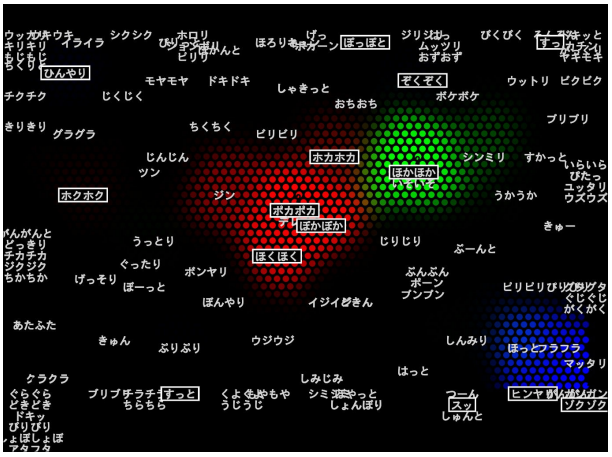

図 6　特定の 3 変数だけに着目した表現（図 4改）

今後は，問題点に挙げた定義に検討を加えるべきであろう．そして，新しい定義に基づいた詳細な実験検討を行っていきたい．

参考文献

[Akita 06] Akita , K : Embodied semantics of Japanese psychomimes, in Proceedings of the Thirtieth Annual Meeting of Kansai Linguistic Society, 関西言語学会, pp. 45-55., (2006).

[Hofmann 99] Hofmann, T.: Probabilistic Latent Semantic Indexing, in Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.50-57, (1999).

[Kohonen 01] Kohonen, T.: Self-Organizing Map, 3rd Edition, Springer-Verlag, (2001). 徳高平蔵，岸田悟，藤村喜久郎訳: 自己組織化マップ, シュプリンガー・ジャパン, (2005).

[工藤 07] 工藤拓，賀沢秀人：Web 日本語 N グラム第1版, 言語資源協会, (2007).

[工藤] 工藤拓: PLSI, http://chasen.org/~taku/software/plsi/

[工藤] 工藤拓：形態素解析器 MeCab, http://chasen.org/~taku/software/mecab/.

[Kurosawa 10] Kurosawa, Y., Mera, K., and Takezawa, T. : Psychomime Classification and Visualization Using a Self-Organizing Map for Implementing Emotional Spoken Dialog System, In Spoken Dialogue Systems Technology and Design, Wolfgang Minker, W., Lee, G. G., Nakamura, S., and Mariani, J. (eds), pp.107-134, Springer-Verlag, (2010).

[黒澤 10] 黒澤義明，目良和也，竹澤寿幸：自己組織化マップ SOM による心情を表すオノマトペ分類の再検討，言語処理学会年次大会, (2010)

[黒澤 08] 黒澤義明，原章，市村匠：換喩検出を目的とした自己組織化マップ SOM による物体の形状マップ生成，言葉と認知のメカニズム，pp.353-374，ひつじ書房, (2008).

[小野 07] 小野正弘：擬音語・擬態語 4500 日本語オノマトペ辞典，小学館，(2007).

[中村 09] 中村沙織，黒澤義明，竹澤寿幸：自己組織化マップ SOM による心情を表すオノマトペの意味分類と可視化，言語処理学会年次大会, (2009)